Evaluating Autonomous Ground-Robots

Anthony Finn[1], Adam Jacoff[2], Mike Del Rose[3], Bob Kania[3], Udam Silva[4] and Jon Bornstein[5]

Abstract

The robotics community benefits from common test methods and metrics of performance to focus their research. As a result, many performance tests, competitions, demonstrations and analyses have been devised to measure the autonomy, intelligence, and overall effectiveness of robots. These range from robot soccer (football) to measuring the performance of a robot in computer simulations. However, many resultant designs are narrowly focused or optimised against the specific tasks under consideration. In the Multi-Autonomous Ground-robotic International Challenge (MAGIC) 2010 the need to transition the technology beyond the laboratory and into contexts for which it had not specifically been designed or tested meant that a performance evaluation scheme was needed that avoided domain-specific tests. However, the scheme still had to retain the capacity to deliver an impartial, consistent, objective and evidence-based assessment that rewarded individual and multi-vehicle autonomy. It was also important to maximise the understanding and outcomes for technologists, sponsors and potential users gained through after-action review. The need for real-time, simultaneous and continuous tracking of multiple interacting entities in an urban environment and over 250,000 square metres in real time compounded the complexity of the task. This paper describes the scheme used to progressively down-select and finally rank the teams competing in this complex and 'operationally realistic' challenge.

Keywords

Evaluation, Ground Robotics, Multi-Vehicle Autonomy, Human-Robot Interaction

[1] Defence & Systems Institute (DASI), University of South Australia, Mawson Lakes, SA 5095
[2] National Institute for Standards and Technology (NIST), Intelligent Sys Div, Gaithersburg, MD 20899-8230
[3] Tank Automotive Research Development & Engineering Centre (TARDEC), Warren, MI 48397-5000
[4] Communications, Electronics, Research Development & Engineering Centre (CERDEC), Ft Monmouth , NJ
[5] Army Research Laboratory (ARL), Powder Mill Road, Adelphi, MD 20783-1197

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
14 JUN 2012	**Journal Article**	**03-01-2012 to 09-05-2012**

4. TITLE AND SUBTITLE
Evaluating Autonomous Ground-Robots

5a. CONTRACT NUMBER
5b. GRANT NUMBER
5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Anthony Finn; Adam Jacoff; Mike Del Rose; Bob Kania; Udam Silva

5d. PROJECT NUMBER
5e. TASK NUMBER
5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army TARDEC,6501 East Eleven Mile Rd,Warren,Mi,48397-5000

8. PERFORMING ORGANIZATION REPORT NUMBER
#22994

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army TARDEC, 6501 East Eleven Mile Rd, Warren, Mi, 48397-5000

10. SPONSOR/MONITOR'S ACRONYM(S)
TARDEC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
#22994

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The robotics community benefits from common test methods and metrics of performance to focus their research. As a result, many performance tests, competitions, demonstrations and analyses have been devised to measure the autonomy, intelligence, and overall effectiveness of robots. These range from robot soccer (football) to measuring the performance of a robot in computer simulations. However, many resultant designs are narrowly focused or optimised against the specific tasks under consideration. In the Multi-Autonomous Ground-robotic International Challenge (MAGIC) 2010 the need to transition the technology beyond the laboratory and into contexts for which it had not specifically been designed or tested meant that a performance evaluation scheme was needed that avoided domain-specific tests. However, the scheme still had to retain the capacity to deliver an impartial, consistent, objective and evidence-based assessment that rewarded individual and multi-vehicle autonomy. It was also important to maximise the understanding and outcomes for technologists, sponsors and potential users gained through after-action review. The need for real-time, simultaneous and continuous tracking of multiple interacting entities in an urban environment and over 250,000 square metres in real time compounded the complexity of the task. This paper describes the scheme used to progressively down-select and finally rank the teams competing in this complex and ?operationally realistic? challenge.

15. SUBJECT TERMS
Evaluation, Ground Robotics, Multi-Vehicle Autonomy, Human-Robot Interaction

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Public Release**	**25**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Introduction

The Multi-Autonomous Ground-robotic International Challenge (MAGIC) 2010 was an initiative sponsored jointly by the US and Australian Departments of Defence in which research awards of $750,000, $250,000 and $100,000 were provided to the top three finalists. The event, was designed to demonstrate emerging unmanned technologies necessary to meet a range of current and future requirements, and was open to national and international organisations within industry and academia. It was conducted in two phases: a two-step down selection to five teams; followed by a multi-robot challenge that featured multiple autonomous unmanned vehicles conducting coordinated intelligence, surveillance and reconnaissance (ISR) missions in a mock urban environment.

The challenge took place during November 2010 at the Adelaide Showgrounds in South Australia. In early October 2009, following an appraisal of their proposals, ten teams were provided with $50,000 research awards for the maturation of their technology; a further two teams were permitted to mature and demonstrate their technology but did not receive funding. Of these twelve teams, following a demonstration of their progress against set tasks, the best five received a further $50,000 award in June 2010. These teams competed in Adelaide.

The challenge was designed to test the ability of multi-robot systems to autonomously and dynamically coordinate, plan and execute an ISR mission in a dynamic environment while simultaneously requiring the provision of a unified situational awareness picture to users removed from this environment. It attracted and energised a wide, international community of academic and industrial participants who brought fresh insights to the problem of identifying and developing robust autonomous multi-robot cooperatives. The rules were complex as the intention was – as closely as possible – to reproduce a realistic ‘operational’ scenario while constraining the magnitude of the design and engineering task to key aspects of multi-vehicle autonomy and control (i.e. aspects of sensing and mobility were avoided).

Each team was allowed a maximum of two operators to supervise a minimum of three ground robots. They were allowed 3 hours, 30 minutes to map the indoor and outdoor environments completely and accurately while detecting, locating, classifying, tracking, and – where appropriate – ‘neutralising’ mobile and static Objects Of Interest (OOI), simultaneously discriminating mobile OOI from non-combatants (neutralising the OOI and/or robots refers to their effective destruction, which was simulated by remotely affecting a system ‘freeze’ that eliminated the OOI (or robot) from taking any further part in that Phase).

The locations and trajectories of the OOI were not known to teams in advance. Mobile OOI could be hostile (humans in red jump-suits) or non-combatant (blue jump suits) and manoeuvred outdoors in scripted patterns at speeds of up to 6km/hr. The mobile OOI could stop, turn, about face, reverse or continue manoeuvring. They could also be located inside buildings, but when inside remained stationary. All hostile OOI had the potential to ‘destroy’ a team’s vehicles if the robots entered their lethality zones (5m radius), the effects of which were constrained by buildings. Static OOI (red bins) could also destroy robots entering their activation zones (2.5m radius) to a maximum range of 10m.

Each robot cooperative had to contain two categories of vehicle: Sensor Robots, which could explore and map the area and detect OOI, and Disruptor Robots, which had the ability to neutralise static OOI, and to which teams were limited to a maximum of three. Disruptor Robots neutralised static OOI by approaching them to within 4.5-2.5m and fixing an eye-safe laser on the OOI for a period of 30sec. The Disruptor Robots were not permitted to share their mapping information with other robots and the Disruptor and Sensor functions could not be transferred between robots. Mobile OOI were neutralised through the simultaneous, coordinated, continuous, and confirmed tracking of the OOI by two Sensor Robots for a period of 15sec. All robots had to be autonomous, weigh less than 40kg (including fuel) and travel at a maximum speed of 10 km/hr.

During the challenge, the location of all detectable mobile OOI and non-combatants outside buildings were provided to teams via a real time data feed simulating the moving target indicator (MTI) provided by an Unmanned Aerial Vehicle (UAV). The MTI feed did not discriminate between hostile and non-combatant OOI and was subject to occlusion by building eves, awnings, trees, and other similar obstacles. Teams were also required to identify and correctly locate and encode into their final GEOTIF maps a range of items known as 'lexicons' (toxic waste barrels, vehicles and open doorways) that were dispersed throughout the site.

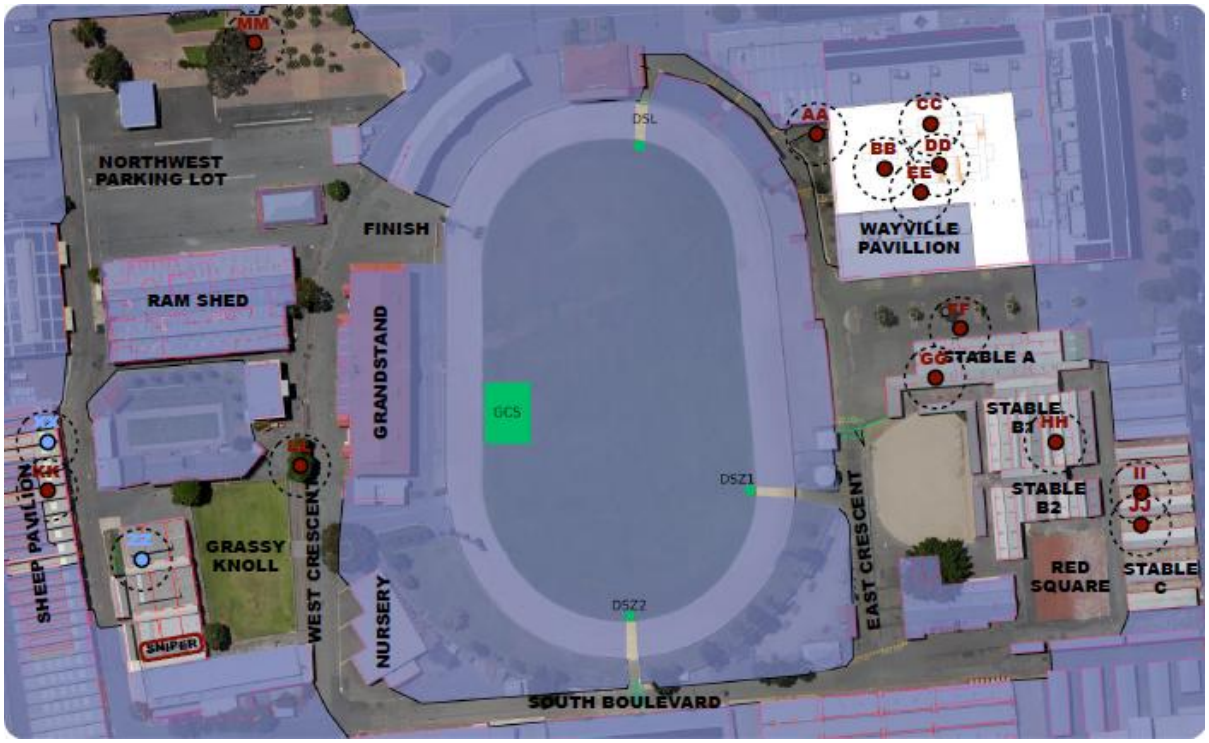

Figure 1: MAGIC 2010 Competition Site, Adelaide Showgrounds, South Australia. The shaded area is 'out of play', the red dots represent the locations of Static OOI, and the blue dots represent the locations of stationary (i.e. indoor) Mobile OOI

The urban environment (Figure 1) covered an area approximately 500m x 500m, which was separated into three increasingly complex phases. Each comprised a combination of roads, paths, buildings, trees, grassy and sandy areas, shallow trenches, safety barriers, curbs and

other obstacles. Buildings were either single or double-storey and of brick construction, with tiled or tin roofs. Teams were provided a representative layout of the area depicting open, restricted and wooded areas, street and road layout, general topography, the location, number, and area of accessible buildings, etc prior to the challenge. However the precise coordinates of the area were only made available to teams as an accurately surveyed aerial image two weeks prior to the challenge. Details such as which buildings and areas were 'in play' and how many OOI there were in Phases I and II were provided to each team the night before they competed.

Phase I of the challenge involved navigation through a complex maze (about 100m x 50m) inside a large building. There were no mobile OOI and the area in play was about 21,000m^2. Phase II required the robots to deal with much more complex terrain for which GPS was partially obscured by urban corridors and subject to significant multi-path effects. There were two mobile OOI (one partially obscured to the UAV feed) but no non-combatants. The area in-play was about 22,000m^2. Phase III covered an area of about 49,000m^2 and involved a total of ten OOI, seven of which were mobile; there were four non-combatants. Phase III also included a sniper, which was undetectable to the robots but which had a similar effect as a mobile OOI (i.e. a single robot was destroyed upon entering the set sniper area). The paths of the multiple mobile OOI intersected in complex and carefully timed sequences. The inside of the buildings in Phases II & III were animal stalls, stables and sheep pens.

Line of sight between the operators and the vehicles was intentionally not possible at all times and the teams were told to consider communications relay. The terrain undulated slightly, but was essentially flat. The ground robots did not encounter significant positive or negative obstacles inside or outside any of the buildings and did not require special mobility or manipulation properties. Access to buildings was through open doorways at least 0.9m wide. GPS was available outside buildings, but was subject to the usual physical restrictions imposed by urban infrastructure. GPS was not relayed inside any buildings.

To circumvent the need for designers to provide endurance to their robots in excess of three and a half hours, teams were permitted to 'service' them (e.g. power up, re-attach cables, electronic re-boot, replace faulty payload modules, re-charge or replace energy sources) at the completion of each phase of the challenge. This could only be undertaken within designated servicing zones and under the supervision of judges. Only minor 'servicing' was allowed (e.g. no changes in software, no replacement of computer boards, no track or wheel changes). Each team was also permitted to walk the course the day before it competed.

The systems engineering and robotic design challenges for participants were numerous and fairly self-evident. Given the complexity and developing nature of the science in a number of key areas, however (e.g. map correspondence, human-robot interaction, effective autonomous robot coordination and behaviour), and the absence of a standard and/or tractable way to assess and distil the many potential metrics into a manageable schema, the challenge for organisers was also non-trivial. They had to establish an objective and fair framework by which increased autonomy, robot-to-robot collaboration, and reduced operator workload could be rewarded and the best teams selected, both during the challenge itself and during the

down-selection process. They also had to monitor the movement and interaction between all of the entities taking part in this challenge in real time. This paper describes the infrastructure and scheme developed by the MAGIC Technology Team to conduct this evaluation.

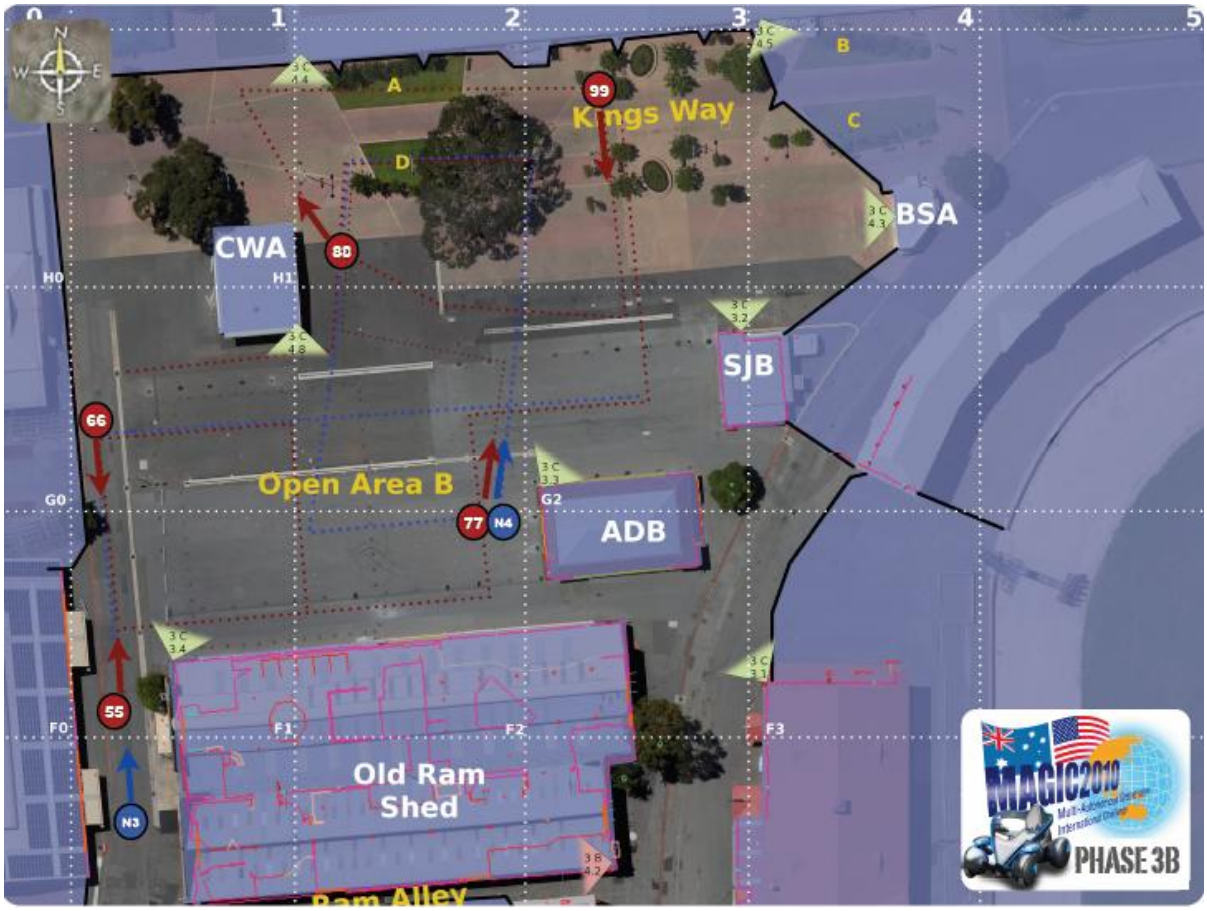

Figure 2: Detail from the later stages of Phase III showing the intersecting paths taken by Mobile OOI and Non-Combatants. The paths and motion were timed and choreographed to make neutralisation possible but difficult. The shaded area is 'out of play'

MAGIC Evaluation Scheme

In addition to the fair placement of teams against one another in the challenge, it was also important to maximise the understanding and the outcomes gained from any analysis. In this regard, there were several other 'clients' for the outcomes of the evaluation: developers and technologists, sponsors of the development work (who are often executive decision-makers or capability planners) and potential users. Consequently, quantitative feedback was needed that each could use to make judgements relative to areas of their work that may need more attention or to allow what would otherwise be an 'apples to oranges' comparison. It was seen as particularly important that these clients understand when such systems will best support their future operations (and perhaps more importantly, when and why they will not). At the outset, therefore, it was agreed that the assessments were to be impartial, consistent across teams and evidence-based, with the evaluation scheme to allocate rewards objectively wherever possible. Because of the complexity, however, a decision was taken to use a few

subjective criteria, but with strict guidelines for the relative qualitative assessments to ensure consistency in ranking across the teams.

The evaluation had to meet several criteria: the analysis had to focus on issues that were relevant to all clients; it had to highlight 'operational' aspects that required more detailed consideration; the outcomes had to be conveyed in language that any decision-maker could understand; and, it had to provide a complete, yet concise, picture of all of the analyses that may need to be conducted. In this regard, it was decided to assess the performance of the teams against three broad categories of effectiveness: Mission Level (performance of against high-level ISR and neutralisation tasks), Systems Level (performance of human and robot as a system) and Technological (performance of the robot as stand-alone system).

In regard to the achievement of any tasks and progress towards them, it was considered preferable to take measurements after the task had been completed [17]. This is because progress on sub-tasks can advance and provide the impression that progress is being made towards an overarching goal, without this actually being the case. For example, in the task 'neutralise OOI' with the sub-task 'advance towards it' if the distance between robot and target reduces, the robot provides the appearance of making progress towards its global goal. If, however, an obstacle is discovered that prevents the robot from advancing towards the OOI, or the OOI was mobile and turned and started approaching the robot, then progress towards the global goal is nugatory.

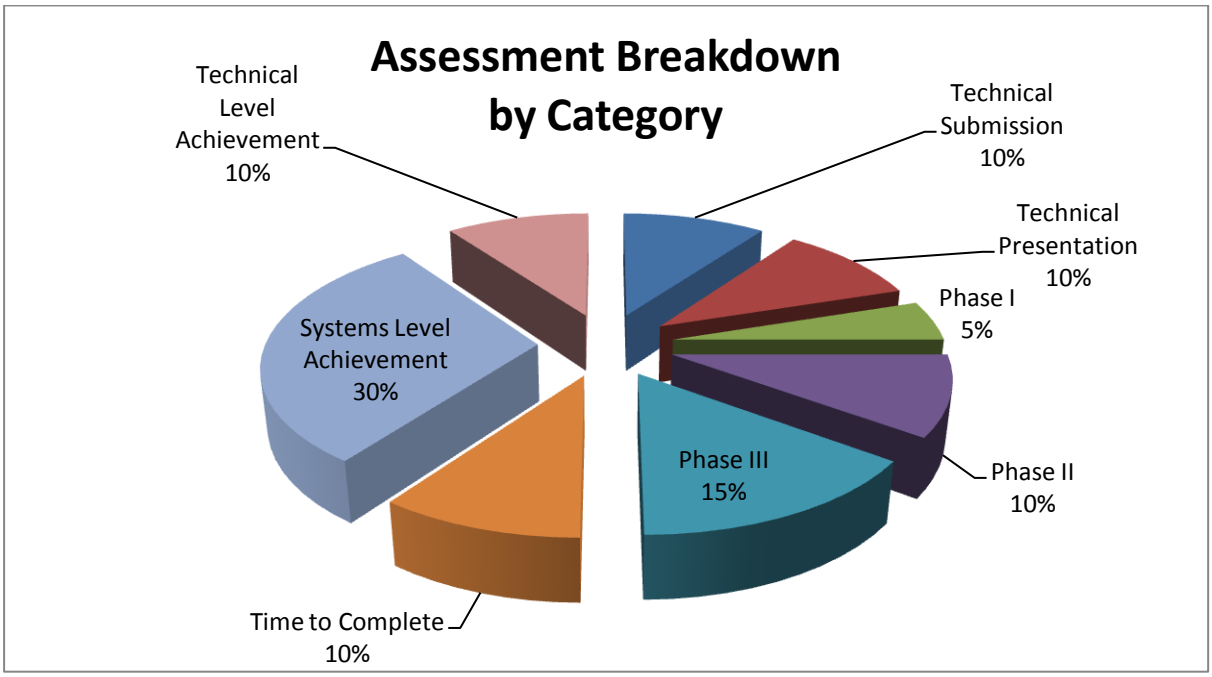

Figure 3: A breakdown of the assessment criteria for MAGIC 2010 vs. category. The Phase I, Phase II, Phase III and Time-to-Complete constitute the Mission Level Achievement criteria.

Early in the challenge planning cycle a decision was made to allocate a total of 1,000 points, of which 200 were to be awarded for a technical report and presentation (100 each) and 800 for the performance of the robots during the challenge. The weighting allocated to mission, system and technical performance criteria was set as 400, 300, and 100 for each of these

criteria, respectively. This reflected the relative balance of technology demonstration and design explanation. The points allocated within each of these broad criteria were then selected against the relative priorities of a small but meaningful number of metrics reflecting the key thrusts and focus of the challenge. Care was taken to also match these metrics against the detailed choreography of the challenge to prevent teams 'gaming' the result (i.e. devising technical solutions that were intentionally optimised against the evaluation criteria without the potential for the technology to transition to a broader military context).

Relationships between the different types of measures are often difficult to establish authoritatively, with linkages between lower-level measures easier to determine than those between higher-level measures which are more dependent on the operational context and hence more scenario dependent. This stressed the necessity for a number of phases (i.e. levels of complexity) within the overall challenge. It was also decided, whenever practical, to ensure that all metrics provided diagnostic information about the dynamics of each phase. In selecting the measures, an attempt was made to meet validity and reliability criteria [10].

The validity criteria required the metric to be:

- Well-defined (measurable data could be collected)
- Relevant (related the technology, system, or mission)
- Realistic (avoided any associated uncertainties)
- Appropriate (related to the key analytical objectives)
- Inclusive (reflected standards required by the analytical objectives)
- Simple (could be easily understood by 'clients')

The reliability criteria required the metric to be:

- Discriminatory (identified differences between alternative designs)
- Flexible (helped to identify new requirements)
- Measurable (could be easily calculated)
- Quantitative (could be assigned a number or rank)
- Auditable (was a clear cause and effect trail)
- Sensitive (reflected changes in system variables)

Mission Level Effectiveness

Complex concepts often require multiple measures to provide valid information. Sadly, no single measure or methodology existed to allow satisfactory assessment of the overall or mission level effectiveness of technology. Consequently, to link the performance of the systems as a whole to the performance of their components, metrics needed to be developed that corresponded to the critical operational and ISR tasks. These tasks required the robots to: accurately and completely explore and map each entire phase area of the site; correctly locating, classifying, recognising and neutralising all OOI in each phase; and completing the entire challenge within the 3hour, 30minute time limit.

Map Assessments

The degree to which teams were able to successfully map the challenge area was evaluated on the basis of *accuracy* and *completeness*. Accuracy was evaluated (in Systems Level Effectiveness – see later) on the extent to which a range of items (the lexicons) dispersed throughout the challenge area were correctly identified and accurately located. The term *accurately located* was defined as the declared location of the observed item touching or overlapping the dimensions of the true/pre-surveyed location of the item or a 1m diameter circle, whichever was larger. Items located within an envelope extending 2m from the edge of the identified item or a 2.5m radius circle centred on the middle of the item, whichever was larger, were considered to be *approximately located* and awarded 'half points'. Items not identified or located outside these envelopes were considered to have been *inaccurately located* and no points were awarded for these items.

Figure 4: Observed map of the inside of a building from a team that would have fared very well for this component (unfortunately, detailed comparisons or results cannot be publicly released). The yellow (1m radius) and blue (2.5m radius) circles show the locations of typical map-evaluation control points (pillars, corners, etc). The red lines indicate paths taken by the robots to obtain these maps.

Completeness, on the other hand, was evaluated on the basis of the local and global spatial consistency displayed between the maps observed by the robots and the ground truth survey data. As there remains no accepted standard for quantitatively measuring the performance of robotic mapping systems against user-defined requirements, and there is no consensus on

what evaluation procedures need to be followed to deduce the performance of such systems, many methodologies currently rely on qualitative analysis (i.e. visual inspection alone). This approach does not allow for better understanding of the specific errors to which the observing systems may be prone and is generally agreed to be sub-optimal. When combined with suitable tools, however, visual inspection techniques do at least provide tractable results in terms of assessing the completeness of an observed map.

To determine the total completeness figure of the area accurately mapped by the robots the site was sub-divided into *(m + n)* smaller local maps of *m* inside areas, $X_{1 \ldots m}$, and *n* outside areas, $Y_{1 \ldots n}$. The observed maps were then superimposed onto maps derived from accurate global ground truth and algorithmically and visually compared to assess the local spatial consistency between the two sets of data. The individual (local) completeness figures, $x_{1 \ldots m}$ and $y_{1 \ldots n}$, were then aggregated into three overall (global) completeness figures using, $A_j = \left(0.25 * \frac{\sum_1^m x_i X_i}{\sum_1^m X_i} + 0.15 * \frac{\sum_1^n y_i Y_i}{\sum_1^n Y_i} \right) * P_j$ where *j* is Phase and P_j = 50, 100, 150.

The software tools that performed these comparisons were developed by the National Institute for Standards and Technology (NIST). They used 'similarity metrics' to measure the quality of the spatial consistency between the observed and true maps, which also gave an indication of the distortion of the map with respect to ground truth. The tools determined the local spatial relationships between key discrete objects (e.g. corners of buildings, doorways, etc) within the observed maps by comparing them to the corresponding objects in the actual environment. On the basis of the number of correct comparisons between the observed and ground truth data sets (relative to the maximum possible), a completeness value was assigned to each local area. The process was then verified by a panel of subject matter experts (SME).

The consistency between the structural detail of the observed maps and those in the real environment was also visually compared to assess the degree to which each map was able to 'hold water' or the structural features of the observed maps suffered from 'noise' (i.e. display jagged, variable or multiple images). Although the effects of noise and distortion between key features did not affect the completeness calculation (i.e. the number of accurately located features in each local map), they were useful in after-action reviews of the data.

In the maze a number of half-barrels were used to bracket several walls. These served as fiducial markers. There is somewhat more to the idea, but this is essentially a low cost way to populate the environment with known references in order to measure the results without accurately geo-referenced ground truth (for programmatic reasons, the maze in Phase I had to be constructed in the days leading up to the challenge event and access to appropriate survey equipment was not an option). The techniques are based on those developed for the RoboCupRescue Robot League, which supports research agendas beyond the robotic capabilities on display at MAGIC 2010. Comparing any 2-D radial displacement of the half-barrels in the observed maps also proved a good measure of how well each maze 'room' was mapped wall-to-wall.

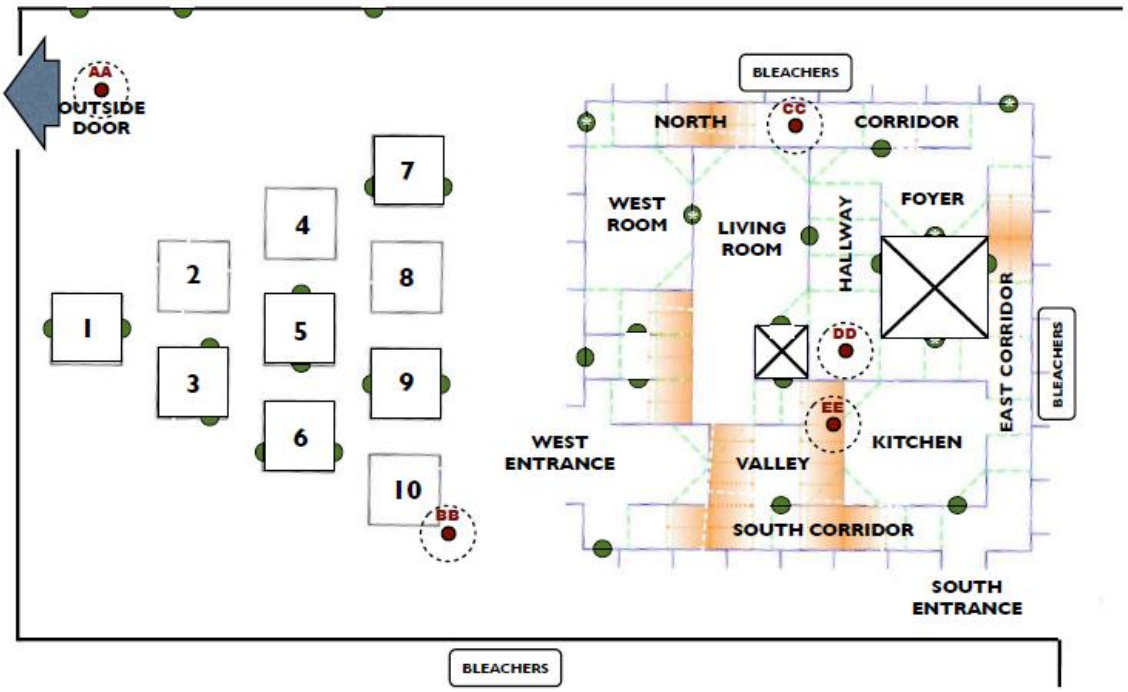

Figure 5: Fiducial markers (half-barrels) located around maze in Phase I

Figure 6: The observed internal structure of the maze in Phase I. The true locations of the barrels are depicted by light green dots; their locations (as determined by the robots) are visible as 'bumps' in the observed data; and the locations of three static OOI are shown as red dots. The mauve lines show the paths taken by the robots. The higher resolution image (inset lower left) shows the effectiveness of the fiducial 'barrel-referencing' technique more clearly.

Taken from the map-assessment tool used to judge accuracy and completeness, Figure 6 shows the fused result of maps generated by one of the team's robot cooperatives. It shows the overlay of the observed and 'approximate' ground truth data. The faint, lighter blues show the ground truth and the heavier, dark blue lines the observed mapping data. The red circles represent the true location of static OOI and the light green circles the true location of half barrels (which obviously coincide with the ground truth). The mauve lines show the paths taken by the multiple robots. The outer dimensions of the maze are about 25m x 25m.

At the level of resolution shown in the main image (higher resolution is available on screen), the outlines of the half-barrels are just visible on the numbered 4m boxes and some internal walls of the maze. Note also the correspondence between the observed 'global' distortions in the map and mismatch between the observed and true locations of the fiducials, as well as the multiple and missing walls on the left of the image. This makes the fiducial barrels easy to count when they are well formed (Figure 6, inset image), which allows easy computation of the completeness metric. When the fiducials are separated due to errors in the robot's pose estimate they break apart in one of two axes. This also permits use as a coarse but obvious measure of performance (e.g. a 'consistency' or 'noise' metric), although this was not employed in MAGIC.

Robot Interaction with OOI

In determining the proportion of OOI correctly identified and located, equal weight was given to the correct identification and location of mobile and static OOI both inside and outside buildings. So, $B_j = MAX\left[0.20 * \left(\frac{M_j + N_j - I_j}{m_j + n_j}\right) * P_j, 0.0\right]$ where N_j is the number of static OOI (out of a possible n_j static OOI) correctly identified and located, M_j was the number of mobile OOI correctly identified and located (out of a possible m_j mobile OOI), and I_j was the number of OOI incorrectly identified. OOI were *correctly identified* and located when declared within a 1m radius centred on the true location of a legitimate OOI, not otherwise correctly identified and located (i.e. M_j or N_j increment, but I_j does not). OOI were *incorrectly identified* when declared outside a 3m radius centred on the true location of a legitimate OOI, not otherwise correctly identified and located (i.e. M_j and N_j did not increment, but I_j did). An OOI declared between 1m and 3m from the true location of a legitimate OOI was deemed to be *correctly identified* but *incorrectly located* (i.e. M_j and N_j did not increment; nor did I_j). In situations where multiple OOI were declared within the 'correct' or 'half correct' zones of legitimate OOI (but where fewer OOI exist), I_i was incremented by the number of 'superfluous' OOI declared.

The weighting for static and mobile OOI neutralization inside and outside buildings was: 15% for static OOI inside buildings regardless whether they are humanoid or static OOI, 10% for static OOI outside buildings, and 15% for neutralisation of mobile OOI (all outside). In other words, if a team neutralised L_j static OOI out of a possible l_j outside buildings, M_j static

OOI out of a possible m_j inside buildings, and N_j mobile OOI out of a possible n_j, they would score $C_j = \left(0.10 * \frac{L_j}{l_j} + 0.15 * \frac{M_j}{m_j} + 0.15 * \frac{N_j}{n_j} \right) * P_j$,

Points were awarded for completing the challenge within T_c minutes in accordance with $\mathrm{D} = 100 * \left(1 - \frac{\mathrm{T_C}}{210} \right) * \frac{1}{3} \left(\frac{A_1}{20} + \frac{A_2}{40} + \frac{A_3}{60} \right)$.

In tandem with the problem of determining the relative value of locating and neutralising mobile and static OOI inside and outside buildings was the requirement to accurately and simultaneously track all OOI (to provide a filtered view for the UAV MTI feed) and ascertain their locations relative to all the robots in real time (e.g. for robot neutralisation purposes). An added constraint was that accurate navigation information could not be interpretable or available to teams. Carried out correctly, quantitative performance data referenced to ground truth positions and time would also be captured as this would assist after-action reviews.

Figure 7: (Lower Left) Asset tracking system components (L to R): ultra-wideband radio frequency receiver (shown with integrated high-gain antenna), 1W transmitter tag & 30 mW transmitter tag. The asset tags transmit a unique identification at up to 50 Hz. Not shown is the network hub, computer, and cabling. (Right) Several badge tags attached to helmets to track personnel in the scenario. (Upper Left) Accurately located barrel construction used to elevate the UWB radio receivers. The black barrels shown were coloured yellow to double as fiducial references for the outdoor map assessment. All cabling was raised above 2m.

Fortunately, NIST and ARL researchers have been working with an asset tracking system, developed by Multi-Spectral Solutions Inc., that captures 2-D location and path data for robots, vehicles, and personnel operating within training scenarios set up to evaluate robotic perception systems. The tracking system relied on state-of-the-art ultra wideband (UWB) radio receivers distributed throughout challenge site to track all dynamic targets (all mobile OOI and robots carried badge-size UWB transmitters). System characterization tests had previously determined the best possible 2-D accuracy of the system as approximately 15 cm (6 inches), although the presence of multi-path reduced this to around 30cm in Adelaide. Unfortunately, previously, the system had only used it to track vehicles and personnel throughout an area of about 80,000 square meters (the MAGIC 2010 site was 250,000 square meters). The system was also used to track the robots through the plywood walls of the maze, achieving similar accuracies. The system could not track through brick or concrete walls. The largest number of dynamic and static transmitter tags used simultaneously was 25 dynamic tags and in excess of 50 receivers, significantly more than had been previously attempted.

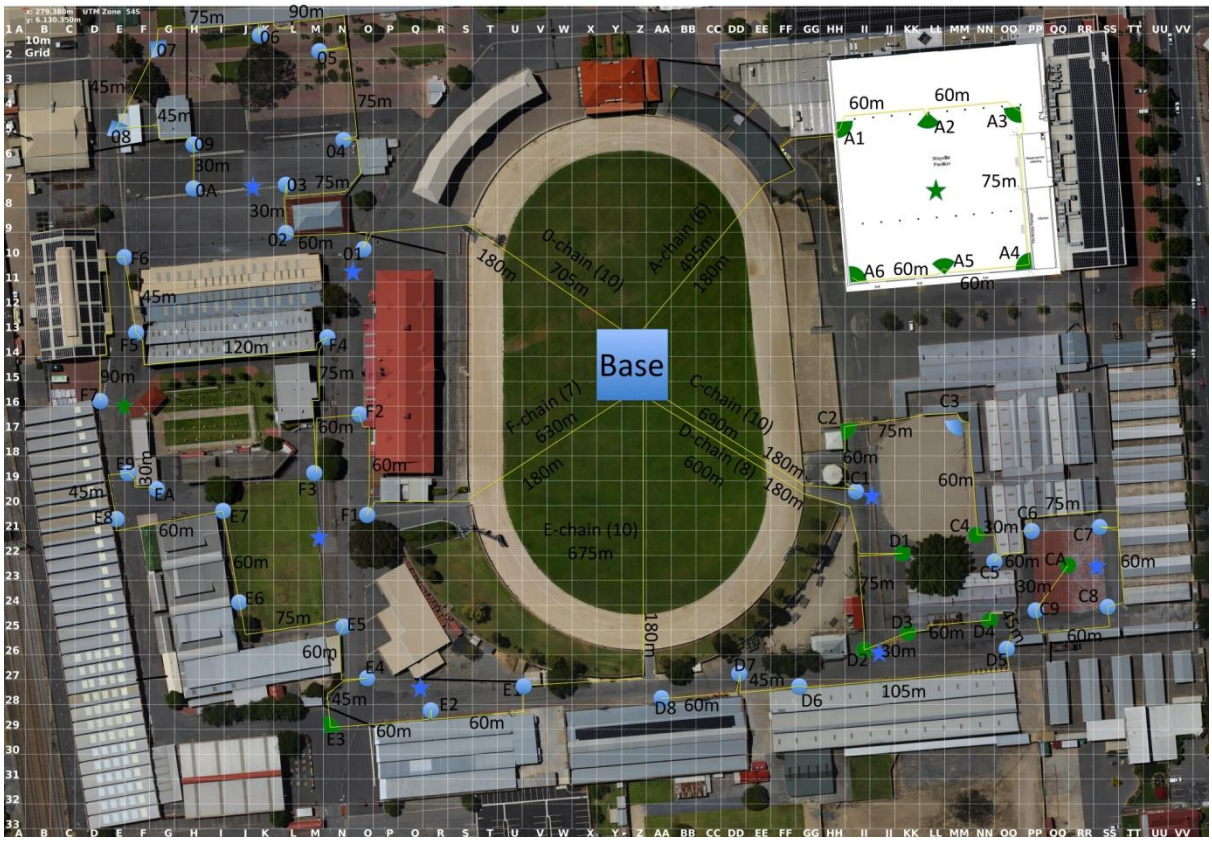

Figure 8: Overview of MAGIC 2010 UWB multi-robot/OOI tracking infrastructure that doubled as fiducial references for the map assessment (blue dots represent barrels holding omni-directional antennas and green cones represent barrels holding directional antennas).

To ensure repeatability all mobile OOI were provided with ‘walker paths’ (faint chalk lines marked at 30m intervals) so that they could accurately and consistently choreograph their movements in accordance with a carefully scripted and inter-dependent mobile OOI scenario.

Around sixty cameras were also deployed to provide judges with multiple perspectives of the robots as they negotiated the site. A number of judges (wearing camouflage; some mobile some stationary) were also discretely located throughout the site.

Although captured within the Systems Level Effectiveness (because the identification and meta-coding process was carried out by the human operators in conjunction with the robots), a total of Q points were awarded according to $Q = 100 * \left(\frac{M + 0.5 * N}{m} \right)$, where M and N were the number of accurately located and approximately located 'lexicons', respectively, out of a total of m possible items (*accurately located* was as defined above).

Systems Level Achievement

Key to the success of military robots is their ability to work as partners with their human supervisors, leveraging the most useful capabilities of each. Basic mission tasks, regardless of the environment, will demand close collaboration. However, because cost pressures and the need to minimise operational risk will keep operator teams small, the effectiveness of the interactions will have a major impact on the effectiveness and performance of future robotic missions. To assess the systems performance of the teams in MAGIC we therefore elected to measure how well the humans and the UVS performed as a team where these interactions included mission planning, plan and task execution and monitoring, plan and problem diagnosis, and the authorisation of mission and plan execution.

There are a number of well-known techniques for evaluating human-robotic teams (HRT) and human-robot interactions (HRI) (e.g. [8] [13] [18] [22] [26] [28]). However, there is not yet a consensus on a standard framework. Furthermore, it is usually possible to carry out tasks using some combination of humans and robot with the key issue being to achieve an optimal mix of automation for each task. Some methods for assessing systems performance are based on decomposing scenarios into 'functional primitives' [27], allocating these primitives either to the supervisors or the automation, evaluating execution of each primitive and computing the ratio of the performance benefit resource allocation. Other methods [24] measure interactive effort; that is, degree of autonomy and/or the overall effort required by the human to work as a component of the supervisor-robot team. This technique is particularly effective when the overall task requires a mix of competencies within the team. However, the effort needed from the human component is a function of the complexity of the task, the circumstances under which it is performed, the training and skills of the human, the perception available through the 'window' of the HMI, the level of trust and confidence enjoyed, and the degree automation imbued within the robot.

As a result, many of the techniques provide a misleading insight as their metrics tend to focus more on the human work component or the UVS component [6]. Also, whereas robots of duplicate design offer repeatable performance, many of the above degrees of freedom vary across individual humans. Additionally, many assessment techniques focus on single robot problems, where the discrete levels of autonomy allow direct comparisons of the system's overall performance to be made against one another. When there are networks of UVS the problem becomes more complex as it is necessary to assess the coupled impact of the levels

of human-UVS automation, the effects of various levels of collaboration between the UVS, the indirect influences of interaction between the automation schemes, and the impact of mission and/or environmental complexity.

Nevertheless, techniques do exist to identify the decision-making roles in which supervisors are most influential and effective relative to the capabilities of UVS [3]. For example, observations can be made under varying levels and types of human intervention and the speed and accuracy of decisions and actions, the time to respond to critical events, the duration of tasks/mission activities, and the ratio for completion of mission-critical vs. secondary objectives all used to estimate the operator-to-vehicle ratios for any given task/mission [22]. This can also be drawn out from the speed and accuracy of the task completion for different levels of task demands associated with the mission (e.g. the number and rate of the required tasks for successful mission completion; the complexity of the mission; etc). These objective measures can then be used to identify (say) the point at which the operators start to shed other tasks or fail to achieve accurate task completion.

MAGIC recognised that the participation of humans at some level was inevitable (and as it turned out, in many cases, continuous). Consequently, the goals of the challenge and the evaluation scheme aimed to reward teams that developed and fielded solutions that were successful in reducing the cognitive and physical workloads imposed on the human operators in favour of these becoming 'machine operation', thereby freeing the humans to perform more complex mental manipulations. This reflected one of the core aims of the challenge that was to shift the state-of-the art relative to the automation of multi-vehicle cooperatives. To measure the amount and frequency of interaction that the individual users had a line was marked on the ground approximately 1m from each team's GCS. Operators then had to stand behind this line such that they were unable to reach their keyboards or displays of their GCS/HMI, without first stepping forward over this line.

In order for either operator to then interact with their robots – even if their interaction was via another modality (e.g. via voice or hand gestures) – an SME monitoring each operator would press START on a chess clock before the user stepped over the line. The operator could then interact with their GCS. On completion of their interactions operators would then step back across the line and the SME would press STOP on their clock (for reasons of accuracy, consistency, and reliability pressure pads would have been preferred, but were unavailable). At the end of the challenge the total interaction time for each operator was calculated and points deducted at 0.2 points per second per operator for all 'over the line' activity.

As the task of 'lexicon' and OOI identification and entry into the map was anticipated as a largely human activity (but could be automated) the 'code' to identify these objects required only a two character manual entry on a standard keyboard. As a result, all teams were be given a 600sec 'interaction allowance' that permitted them to interact with their robots for the purposes of object identification without penalty. The following criteria were then used to measure the effectiveness of the human-robot system.

The sophistication of the HMI and the completeness of its situational awareness displays were assessed subjectively by SME against its capacity to deliver streamlined information of high usability to military operators. For example, its capacity to allow accurate and direct perception, comprehension and prediction of challenge activity and events; its capacity to prompt for or autonomously allocate resources, monitor uncertainty and risk, and thereby provide automated assistance; and/or its capacity to correspond multiple sensory perspectives into a readily interpretable common operating picture were evaluated [4] [11]:

HMI Design & Display Real Estate

- Was all of the key information readily available
- Did the display omit any information
- Was information difficult to access
- Did information require time to retrieve
- Was the information all clearly marked
- Were font sizes & colours appropriate
- Were there moving map displays
- Were the displays/windows easily navigable
- Was information 'buried' deep within a display
- Were appropriate displays/windows located next to each other
- Was a significant amount of cognitive resource required to navigate the displays

Operator Attention

- Were operators distracted from their primary tasks by secondary tasks
- Did individual displays (e.g. 'pop-ups') obscure an operators view of the HMI
- Did the HMI (or did it fail to) alert the operator to any key events
- Did the HMI make use of audio and visual cues/modalities

Cognition & Workload

- Was the operator's mental workload high/low
- Did the operator have to search for information
- Did the operator have to mentally integrate/manipulate information from multiple screens, sensors, etc or was this executed autonomously
- Did the displays support direct perceptual interaction
- Did operators have to 'remember' key information or was this 'embedded' in the system (e.g. speed limits, detection/activation zones, etc)

Change Analysis & Situational Awareness

- Was the information up to date
- Were OOI and robot status reported
- Were fuel/battery life/etc information reported
- Were potential/past OOI paths shown/available
- Were robot options/paths available

- Was a risk analysis presented/available
- How was novelty dealt with

Planning & Execution Tasks

- Did the HMI permit effective and efficient decision-making
- How was event and information cueing delivered wrt sensors, robots & humans
- Did the HMI permit constraint violations (i.e. allow robots into known threat area or provide plans that exceeded battery life)
- Did the HMI attempt to negate any inadequacies in the operators' partial or conflicted understanding of information
- Did the HMI's functions/display vary according to human or robot workload
- Were users prompted (or receive autonomously) feasible resource allocations

Finally, users task a system by entering specific goals, constraints, or mission priorities through the HMI. These tasks might include map exploration, OOI acquisition, maintaining surveillance over a region, move to waypoint X, or respond to static and mobile OOI threat Y. There may also be constraints that are task dependent such as OOI-based no-go zones, predicted communications 'dead-spots', or time constraints – and which teams enter as 'hard' or preference-based options. In this regard, the 'level' and type of planning and the time taken (relative to mission time) were also noted.

The coordinated neutralisation and response to mobile OOI manoeuvres were penalised on the basis of the number of robots lost to mobile OOI action and the number of interruptions to the neutralisation process. Points were awarded if no robots are lost during the mobile OOI neutralisation process (at least 3 mobile OOI had to be successfully neutralised for these points to be awarded) with 20% deducted from this total each time a robot is lost during the mobile OOI neutralisation process. More points were awarded if all of the mobile OOI neutralisation attempts succeeded first time, with 10% deducted for each time a neutralisation attempt was re-started due to the robots losing track of the OOI.

Points were also awarded for the human-to-robot ratios maintained throughout the challenge operations (although only robots actively participating in the challenge were used in determining these ratios). Maximum points were awarded if teams correctly maintained 2:1 sensor-to-disruptor and 2:3 human-to-robot ratios throughout challenge at all times regardless of sniper, mobile and static OOI activity, robot reliability, etc. Teams then lost 20% if the sensor-to-disruptor ratio fell below 2:1 and a further 20% if this ratio fell below 1:1. Teams lost a further 20% if the human-to-robot ratio fell below 2:3 and a further 20% if this ratio fell below 2:2.

The number and nature of tasks/sub-tasks carried out autonomously and collaboratively by the robots were also rewarded according to the subjective discretion of the judging panel:

- Were appropriate multi-robot autonomy and collaboration strategies selected?
- How appropriate were these strategies & how autonomously were they executed?
- How quickly were key decisions made and how 'correct' were they?

- Were these decisions made by the robots, the humans or both?
- How quickly could decisions be interpreted and/or submitted?
- Did the system do what the human expected and was this appropriate to the task?
- How well did the automation/robots handle any unforeseen events and/or errors?
- How much human resolution of ambiguities and uncertainties was required?
- What was the impact of mission re-planning time on mission success?
- Did the loss of assets affect achievement of objectives and/or mission success?

Technical Level Achievement

We also needed to measure the functionality of the system and its components in terms of their technological drivers. To do this systematically, we need metrics that allow evaluation relative the performance of the robot in terms of its efficiency and effectiveness. As many robots will finish tasks if allowed time, however, we also had to measure the number of tasks completed regardless of human intervention versus the number completed autonomously.

There have been several attempts to develop taxonomies for measuring the performance of robots (e.g. [8] [13] [28] [30]). As with measuring HRT/HRI performance, however, there is not yet a consensus on a standard framework. Furthermore, many of the frameworks strive to establish performance against what individual robots were designed to do and hence what constrained their design. Regardless of their design, role, or domain, however, military robots operating in real world conditions have at least two things in common: they are designed to meet a set of requirements that are subject to constraints such as budget, schedule, etc, and they are used in environments that differ from those for which they were designed and tested.

The operational reliability of the robots was rewarded. That is, the number of robots actively participating in the challenge that broke down or had to be disabled (breakdown was defined as suffering from a hardware or software malfunction causing the robot to become unresponsive to commands issued via the HMI; robots that were legitimately serviced under the supervision of judges or that were destroyed (and responded appropriately) were not considered to have broken down. Points were therefore awarded according to the percentage of robots 'in play' that broke down, maximum points being awarded for no breakdowns, etc.

The directness of the robot's routing and burden-sharing was also rewarded on a graded scale of how evenly these tasks were shared between the robots and the variability with which they planned and executed their navigation paths.

Finally, robot automation was rewarded by deducting points on the basis of the number and duration of the interventions made by each team. In addition to 0.2 points per second per intervention per operator being deducted from the total gained in any single phase (negative points was not possible), teams were ranked in accordance with the number of 'over the line' interventions. The ratio of the total intervention time, T_I, to the time taken to complete the challenge, T_c, also resulted in a further $20 * \left(1 - \frac{T_I}{T_c}\right)$ points being awarded.

Finally, teams were penalised for neutralising a non-combatant (all phase points), each robot destroyed by an OOI (10 points per event) and collisions (10 points per event). Collisions with obstacles or infrastructure were determined on the basis of kinetic energy imparted by the robots. For instance, collisions that required intervention from the organisers to rectify the effect of the collision resulted in point deductions. However, if a robot gently bumped (but did not move) an item of infrastructure, no points were deducted. However, if the item moved or if damage was done (say) to a wall or a doorway, points were deducted. All collisions between robots were penalised.

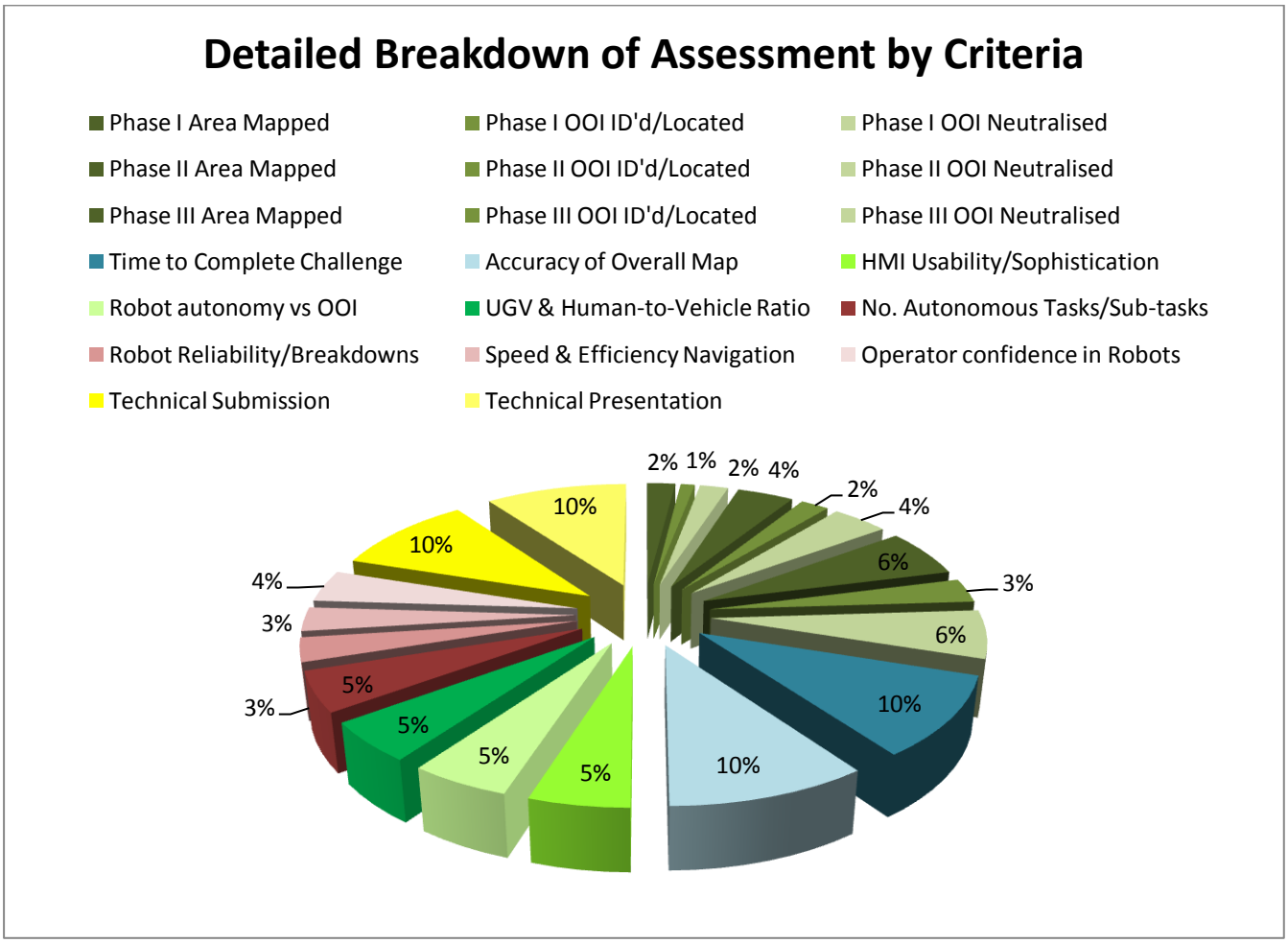

Figure 9: Relative impact of assessment criteria for MAGIC 2010 on scoring outcome

Evaluation during Phased Down-Selection

MAGIC used a two-step down selection to arrive at the five finalists. In early October 2009, following an appraisal of their proposals by a technical panel, ten teams were provided with research awards for the maturation of their technology and a further two permitted to mature and demonstrate their technology without funding. In June 2010, in addition to presenting their designs, the twelve semi-finalist teams then had to demonstrate a range of system and individual robot functionality against set tasks that closely mirrored what they would face in the full challenge. These tasks included: autonomous indoor and outdoor robot navigation; multi-robot supervision/control using the GCS/HMI; autonomous multi-robot collaboration

and correspondence of multiple maps; detection, classification, recognition, identification, and neutralisation of a static OOI; detection, classification, recognition, identification and neutralisation of a mobile OOI (and discrimination from a non-combatant); communications relay/re-transmission (if pertinent); and emergency stop and safety mechanisms.

Teams' performance was then determined against tasks that included: mapping, navigation, situational awareness, HMI & GCS design, multi-vehicle coordination, neutralisation, system and self-awareness, and a raft of other 'miscellaneous' criteria such as usability, engineering design, spectrum management, and operator workload.

Mapping – To determine the completeness of the area accurately mapped by the robots, the observed map was recorded as a standard GEOTIFF and superimposed onto maps derived from accurate global ground truth. These maps were then visually compared to the truth to assess the spatial consistency between the two sets of data broadly in accordance with the techniques outlined in previous sections of this paper.

Navigation – This was considered to be a fundamental task required of an autonomous robot and was measured by how well the robot was able to determine where it was, where it needed to be, how it got from where it was to where it needed to be in terms of its trajectory management and resource usage, and how it dealt with environmental contingencies along the route. This was accomplished by the Technical Assessment Panel (TAP) measuring and noting such things as:

- The success rate in achieving any global goals (completion indoor/outdoor course)
- The efficiency with which navigation was carried out
    - Time to complete an aspect of the course (did robots 'get stuck' in corner)
    - Distance travelled (and time taken) to reach key waypoints
    - Steps required to reach each of these key waypoints
    - Area mapped per robot (and/or per unit time)
- The effectiveness with which obstacle avoidance functions are executed
    - Time taken for computation of hazard detection (did robots 'stop to think')
    - Number and nature of obstacles detected, avoided, etc
    - Number and nature of obstacles that could not be avoided
    - Percentage of navigation tasks completed
    - Range at which obstacles are identified (this was inferred)
    - Deviation from planned routes (extent and nature of deviations)
    - Coverage of area (even distribution of robots, repeated passes of an area)
- Operator confidence in the navigation capabilities
    - Number of operator interventions per unit time (planned or unplanned)
    - Ratio of operator input time to robot navigation time
    - Did robots 'get stuck' in corners or for any reason
- Average and maximum speed over terrain
    - Ratio of average to maximum speed
    - Percentage time spent at maximum speed
    - Total time taken to complete navigation tasks

- Degree to which environment confuses robot's navigation sensors
    - Elevated floors, slippery surfaces, barriers, mesh netting, no GPS, etc.

Situational Awareness – This was determined by measuring the capacity of a robot to:

- Detect items
    - Determination that an object may be of interest (take closer look)
- Classify items
    - Object can be discriminated by class (object is mobile not static)
- Recognise items
    - Object can be distinguished by category (mobile OOI not non-combatant)
- Identify items
    - Object is distinguished by model (object is mobile OOI # 35)

HMI & GCS – These were characterised as per the previous sections of this paper.

Management & Coordination – These tasks required measurement of the vehicle-to-human ratio, the human-robot interaction, systems performance, problem recognition, teaming, degree of heterogeneity and information sharing. Essentially, this was related to the difficulty in handling the robot team during use through the traditional way in which this is measured: through the use of the *fan out, intervention, attention demand,* and *free time* metrics. These criteria measure of how many robots can be effectively controlled by a human and the amount of operator intervention that needs to be devoted to them.

The *fan out* measure relates to the logistical demands of the robots, their handling difficulties and the related cost-benefit equation; for example the number of robots an individual operator is able to effectively supervise. The measure is also a good indicator of intervention required to manage the robot cooperative. The *intervention* metric usually measures the physical or cognitive intervention response time from when the operator first recognises the problem or from when the robot first requests assistance. Response time can also allow for specific details of the intervention or latencies to be measured, such as the time to deliver the request from the robot, time for the user to notice the request, situational awareness or planning times, and execution time. *Attention demand* measures the fraction of the time given by a user to operating the robot, whereas *free time* is a measure of how much time is left over for other tasks. The aim of the teams should clearly be to maximise free time. However, they will likely fill free time with other operations, which means that free time alone is insufficient for characterising efficient human-robot interaction [4].

Neutralisation Tasks - Key measures for these tasks were the ratio of intended to unintended 'victims' (robots or non-combatants) and the physical or cognitive processing requirements placed upon any human operators. For example, the sort of activities that were performed by operators included object identification, association, tracking, and path prediction for mobile OOI. These tasks were measured at an outcome or a 'contact' level by noting the number of correct and inappropriate detections and identifications made and the speed, directness and accuracy with which the overall neutralisation process took place, with the degree of error monitored in terms of major, minor, or inconsequential. Additionally, the coordinated

neutralisation and response to mobile OOI manoeuvres was assessed on the basis of the number of robots lost during the process and the number of interruptions or human interventions required to achieve it.

Miscellaneous – A number of other criteria were also useful [1] [11] [15] [21] [23] [24].

Operator Workload - Was any degree of effort or frustration experienced? If so, what was its impact on mission effectiveness (i.e. was there task shedding)? Was it possible to discern overall & critical component workload profiles? What were the mental, physical, temporal mission demands on the operators?

System Usability - Scales of usability, consistency, reliability were noted. What was the degradation in task effectiveness over robot neglect time? How frequently & meaningfully did robots & user communicate? What was the appropriateness of info exchange between human & robot? What was the timeliness of information exchange between human & robot?

Planning - What was the planning and re-planning frequency? How much and what type of information was required a priori? What was the time from the presentation of task to execution of mission? What percentage of the mission was spent planning (by robot and human)? What was the ratio of operator-initiated input vs. robot-generated prompts? What was the accuracy and completeness of requirements for robot inputs? What was the level of operator assignment (mission and task level)?

Mission Tasking - How long did it take to enter/task each mission, task or sub-task? What types and amounts of information are required for each of these? What were the number and nature of tasks undertaken simultaneously? What were the mission points at which key tasks or sub-tasks were dropped? What was the task completion time vs. required mission completion time? What was the 'accuracy' of task completion vs. task completion time?

Engineering Design – What was the stability and robustness of software, mechanical electrical hardware design? Engineering quality (hardware & software), construction, and supportability? Were there any electromagnetic compatibility/interference issues?

Communications & Spectrum - How appropriate was the communications spectrum usage? What was the quality of communication (user-robot and robot-robot)? What was the complexity of the information flows (user-robot and robot-robot)? Over what range was single/multiple link communications relay demonstrated? To what degree was the simulated UAV feed used/demonstrated?

Multi-robot Tasking - Did complexity or novelty in the environment affect performance? How did the number of robots affect the level of performance? How did the distribution of robots (or OOI) affect performance? What tasks/environments required multi-robot cooperation? What tasks/environments were improved by multi-robot cooperation? What cooperation emerged as a result of interactions between robots? What was the number of tasks that could be handled simultaneously? How many different task or mission types could the system handle? How many tasks were dropped vs. the total number of tasks undertaken?

What proportion of tasks/missions was not achieved (feasible tasks only)? How successfully did the robots fuse/correspond their multiple (local) maps?

Overall Performance – Finally, a number of other subjective considerations that were useful in determining the overall grasp that each team had of the challenge included such things as: Was an appropriate mission strategy selected? How appropriate was the strategy and how well was it executed? How quickly were decisions made and how 'correct' were they? How quickly could any decisions be submitted to the robots? How quickly could these decisions then be interpreted by the robots? How well did the automation handle novelty, unforeseen events, or errors? Did freezing the robot affect achievement of objectives or mission success? Did the system do what the humans expected – and was this appropriate? How many feasible alternative strategies were derived and presented? How varied were each of these (feasible) alternative plans? What was the impact of mission re-planning on mission success? What was the impact of mission re-planning time on mission success? What was the impact of command and re-planning frequency? What was the impact of decision accuracy and error recovery? Was the human taking control when and where appropriate? Was the robot handling its duties when and where appropriate? How much data interpretation was necessary by the human? How much work was needed for the human to direct the automation? How much human resolution of ambiguities and uncertainties was required? Was there agreement between the human and robot perception of the situation? To what extent did the operators accept or reject the robot's operational picture?

Teams were ultimately ranked in accordance with their grasp of the problems and their approach to overcoming them, their quantitative understanding of the key design points, together with the extent to which they offered an optimised design that had demonstrated results. The assessors also looked for an effective systems' engineering approach and test methodology that had been used to evaluate the performance of the existing design and that would minimise future risk.

Concluding Remarks

MAGIC 2010 presented significant challenges to the participating teams in terms of multi-robot design and systems engineering. However, the absence of common test methods and agreed-upon metrics of performance, combined with the need for the assessments to satisfy multiple customers, meant that a performance evaluation scheme was needed that avoided domain-specific tests. This was reinforced by the need for the technology to be transitioned into contexts for which it had not specifically been designed or tested. The scheme developed delivered an impartial, consistent, objective and evidence-based assessment that rewarded individual and multi-vehicle autonomy, while simultaneously maximising the understanding and outcomes for technologists, sponsors and potential users gained through after-action review. The need for real-time, simultaneous and continuous tracking of multiple interacting entities for the purposes of real time adjudication compounded the complexity of the task.

There were, of course, many factors that had the potential to affect team performance that organisers did not include in the final assessments. These included many 'human factors'

issues like team dynamics (the specific roles and relationships of the two operators, how well they worked or didn't work together, how well they knew the rules of the competition, and/or how well they formulated and executed mission or task strategies outside of those posed by their autonomy). Largely, such criteria were omitted for practical reasons such as the capacity to measure them objectively and then quantify their influence on the challenge outcome. There was also the need to have a manageable rubric with which the teams and organisers could ultimately work.

Acknowledgements

The MAGIC 2010 Challenge was sponsored by TARDEC, DSTO, DARPA, ARL, CERDEC, DASA-DEC, DASA-RT, AMRDEC, USAIC, USAITC-PAC, and ONRG. We are very grateful to these organisations and a significant number of our colleagues who assisted in the development and execution of MAGIC 2010 and the development of this evaluation scheme. These include Dr Jim Overholt, Dr Grace Bocheneck, Mr Steve Quinn, Brig (ret'd) Justin Kelly, Brig (ret'd) Steve Dunn, Col Tom Steiner, Lt-Col Eric Stierna, Vinod Puri, Brian Skiba, Nick Lynch, Maryanne Fields, Despina Fillipidis, Bernard Theisen, and Kayla Hahka.

References

[1] Bruemmer, D, et al, *How to Trust Robots Further than we can Throw Them*, Proc. CHI Conference on Human Factors in Computing Systems, Vienna, April 2004

[2] Cohen, M., R. Parasuranman, & J. Freeman, *Trust in decision aids: a model and its training implications,* Tech. Report USAATCOM TR-97-D-4, 1999

[3] Cummings, M.L., *Supervisory Control Challenges in NCW*, University Virginia, Doctoral Dissertation, 2003

[4] Cummings, M.L. et al, *Predicting Operator Capacity for Supervisory Control of Multiple UAVs*, Innovations in Intelligent Machines, Springer, 2009

[5] Davidson, E.J., *Evaluation Methodology: The Nuts & Bolts of Sound Evaluation*, Sage Publications, 2005

[6] Delude, C.M., *MIT Neuroscientists See design flaws in Computer Vision Tests,* McGovern Institute, January 2008

[7] Dixon S.R. & C.D. Wickens, *Control of Multiple UAV's - A Workload Analysis,* Proceedings 12th International Symposium Aviation Psychology, Dayton, OH (2003)

[8] Draper, J.V. & D.B. Kaber, *Human-Robot Interaction,* Industrial & Occupational Ergonomics: Users' Encyclopaedia (A. Mital et al editors), 1999

[9] Endsley, M.R. & D.B. Kaber, *Level of Automation Effects on Performance, Situational Awareness, and Workload in a Dynamic Control Task*, J. Ergonomics, Vol. 42(3), pp462-492, 1999

[10] Finn, R.A. & G. Chalmers, *Operations Analysis Methodology for Force Level Electronic Warfare*, DSTO-RR-0298, DSTO Research Report, 2001.

[11] Finn, R.A. & S. Scheding, *Challenges for Autonomous & Unmanned Vehicles,* Springer, ISBN-978-3-642-10703-0, March, 2010

[12] Goodrich, M. & D. Olsen, *Seven Principles of Efficient Human-Robot Interaction,* Proc IEEE Conf. Systems, man & Cybernetics, p3943-3948, 2003

[13] Granda, T. M. Kirkpatrick, T. Julien, L. Peterson, *The Evolutionary Role of Humans in the Human-Robot System*, Proc. Human factors Society 34th Annual Meeting, p664-668, 1990

[14] Grocholsky, B. *Information-Theoretic Control of Multiple Sensor Platforms*, PhD Thesis, University of Sydney, 2002
[15] Hart, S. & L. Staveland, *Development of the NASA Task Load Index: Research in Empirical and Theoretical Results,* in Human Mental Workload (eds. P. Hancock & N. Meshkati), North Holland Press, 1988
[16] Hinds, P.J., T.L. Roberts, H. Jones, *Whose Job Is It Anyway? A Study of Human-Robot Interaction,* J. Human Computer Interaction, Vol. 19, 2004
[17] Huang, H.M., *Autonomy Levels for Unmanned Systems (ALFUS) Framework,* Proceedings 2006 Performance Metrics for Intelligent Systems (PerMIS) Workshop, Gaithersburg, MD August 21 - 23, 2006
[18] Jacoff, A., Messina, E., Evans, J., *Performance Evaluation of Autonomous Mobile Robots,* Industrial Robot: An International Journal, Volume 29, Number 3, 2002
[19] Kaupp, T. & A. Makarenko, *Measuring Human-Robot Team Effectiveness to Determine an Appropriate Autonomy Level,* Proc. IEEE Conf on Robotics & Automation, Pasadena, CA, USA, May 19-23, 2008
[20] McCarley, J.S. & C.D. Wickens, *Human Factors Concerns in UAV Flight,* University of Illinois HF Division Technical Report AHFD-05-05/FAA-05-1
[21] McCarley J.S. & C.D. Wickens, *Human Factors Implications of UAVs in the National Airspace*, Aviation Human Factors Division, Institute of Aviation, Technical Report AHFD-05-05/FAA-05-01, April, 2005
[22] Nourbakhsh I.R. et al, *Human-Robot Teaming for Search & Rescue,* Pervasive Computing, IEEE Computer Society, January-March, 2005
[23] O'Day, S. et al, *Metrics for Intelligent Autonomy,* Proc. Performance Metrics for Intelligent Systems (PerMIS), 2004
[24] Olsen, S.R. & M.A. Goodrich, *Metrics for Evaluating Human-Robot Interactions,* Proc. Performance Metrics for Intelligent Systems, Gaithersberg, MD, 2003
[25] Rantanen, E. & A. Nunes, *Taxonomies of Measures in Air Traffic Control* Research, Proc. International Symposium on Aviation Psychology, 2003
[26] Rehmann, A.J., *Guide to Human Performance Measurements and Crew Requirements,* Dept transport Technical Report, DOT/FAA/CT-TN95-49, 1995
[27] Rodriguez, G. & C. Weisbin, *A New method to Evaluate human-robot systems performance,* Autonomous Robots, Vol. 14, p165-178, 2003
[28] Steinfeld, A. et al, *Common Metrics for Human-Robot Interaction,* Proc. Human-Robot Interaction 2006, Salt lake City, Utah, March 2006
[29] Wickens, C., *Automation in Air Traffic Control: The Human Performance Issues,* Automation Technology & Human research trends (Eds: M. Scerbo & M. Mouloua), Lawrence Erbaum Assoc., 1999
[30] Yanco, H. & J. Drury, *A Taxonomy for Human-Robot Interaction,* Proceedings of the AAAI Fall Symposium on Human-Robot Interaction, p111-119, 2002